Recipe for Uncovering Predictive Genes Using Support Vector Machines Based on Model Population Analysis

Hong-Dong Li, Yi-Zeng Liang, Qing-Song Xu, Dong-Sheng Cao, Bin-Bin Tan,
Bai-Chuan Deng, and Chen-Chen Lin

Abstract—Selecting a small number of informative genes for microarray-based tumor classification is central to cancer prediction and treatment. Based on model population analysis, here we present a new approach, called Margin Influence Analysis (MIA), designed to work with support vector machines (SVM) for selecting informative genes. The rationale for performing margin influence analysis lies in the fact that the margin of support vector machines is an important factor which underlies the generalization performance of SVM models. Briefly, MIA could reveal genes which have statistically significant influence on the margin by using Mann-Whitney U test. The reason for using the Mann-Whitney U test rather than two-sample t test is that Mann-Whitney U test is a nonparametric test method without any distribution-related assumptions and is also a robust method. Using two publicly available cancerous microarray data sets, it is demonstrated that MIA could typically select a small number of margin-influencing genes and further achieves comparable classification accuracy compared to those reported in the literature. The distinguished features and outstanding performance may make MIA a good alternative for gene selection of high dimensional microarray data. (The source code in MATLAB with GNU General Public License Version 2.0 is freely available at http://code.google.com/p/mia2009/).

Index Terms—Informative gene selection, cancer classification, support vector machines, margin, model population analysis.

1 Introduction

The developed microarray allows scientists to monitor expression levels of thousands of genes associated with different diseases in a very quick and efficient manner. In combination with bioinformatics data analysis methods, such technologies have been gaining extensive applications in the field of cancer classification, aiming at first uncovering the genetic cause that underlies the development of many kinds of human diseases [1], [2], [3], [4], [5], [6], [7] and then administering an appropriate therapy to the patients. Up to date, microarray based cancer classification has acquired a critical role in cancer treatment related areas and the study of cancer classification using gene expression profiles has been reported in an amount of literature [1], [2], [3], [4], [5].

However, the number of genes resulting from microarray experiments is in most cases very large. In contrast, the number of tissue samples is very small. This setting makes the prediction of the tissue phenotype a challenging "$\mathrm{large}\ p, \mathrm{small}\ n$" problem [6], [7]. Moreover, the disease relevant genes usually occupy only a small percent, making it difficult to identify them from the large pool of candidates. However, from the point of view of clinical practice, it is important to identify a small number of informative genes for thorough understanding of the pathogenesis and accurate prediction of clinical outcomes [8]. For this reason, many variable selection methods have been proposed or applied to seek the potential genes which are responsible for tissue phenotypes. Golub et al. proposed to use class distinction correlation for screening the potential gene markers and suggested a general strategy for discovering and predicting cancer [9]. Ma and Huang [10] developed a novel approach for biomarker selection by using the ROC technique with applications to microarray data. In their method, a sigmoid approximation to the area under ROC curve is proposed as the objective function for classification. Their approach proved to yield parsimonious models with good predictive performance. Ghosh and Chinnaiyan [11] performed gene selection and cancer classification by using LASSO [12], which is a widely used method for automatic variable selection and model building. By employing normalized mutual information, Liu et al. [13] presented an entropy-based iterative algorithm for selecting a subset of genes with maximal relevance and minimal redundancy.

Although variable selection methods have been shown to be useful in revealing disease relevant genes, they have one obvious weak point which in our opinion should be addressed and can be improved. The weak point is that the influence of sample variation is not taken into account by the current variable selection methods, indicating that some "bad" variables may be selected as "good" ones by chance (false positives). For example, given a training set, LASSO can output a fixed variable rank. If we remove some (10 percent) samples (causing sample variation) from the training set, the variable rank by LASSO maybe change a

- *H.-D. Li, Y.-Z. Liang, D.-S. Cao, B.-B. Tan, B.-C. Deng, and C.-C. Lin are with the Research Center of Modernization of Traditional Chinese Medicines, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, P.R. China.*
  E-mail: lhdcsu@gmail.com, yizeng_liang@263.net, oriental-cds@163.com, tbb0301@yahoo.com.cn, dengbaichuancsu@163.com, kittycc932@sina.com.
- *Q.-S. Xu is with the School of Mathematical Sciences, Central South University, Changsha 410083, P.R. China. E-mail: dasongxu@gmail.com.*

Manuscript received 22 Jan. 2010; revised 20 May 2010; accepted 14 Oct. 2010; published online 17 Feb. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2010-01-0033.
Digital Object Identifier no. 10.1109/TCBB.2011.36.

1545-5963/11/$26.00 © 2011 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

lot. In order to overcome this problem, we in the present work focus on establishing a variable selection method which takes into account the sample variation and can uncover statistically significant variables. Also, this method is specially designed to work with support vector machines. Reasons for this are: 1) that there are few papers addressing variable selection for SVM [14], [15], [16] and 2) that variable selection should benefit the predictive performance and the interpretability of a SVM model [17]. Guyon et al. utilized the recursive feature elimination (RFE) strategy, which starts from a SVM model built on all the variables and eliminates the variable in a recursive manner, to rank nested subsets of variables according to the weight value of the SVM classifier [17] and greatly improved the performance of SVM. Gualdrón et al. recently proposed a method for variable selection for SVM. The variables are ranked based on the absolute changes of margin of SVMs after only one variable is removed [18]. It is illustrated that better predictive ability is achieved compared to that of using all variables. Recently, Aksu et al. demonstrated that RFE objective function is not generally consistent with the margin maximization principle thus proposing an explicit margin-based feature elimination (MFE) for variable selection of SVMs. They showed that MFE could improve both margin and generalization accuracy [16], [19].

The method reported here, named margin influence analysis (MIA), is quite different from previous work. it is developed based model population analysis (MPA), which is a general framework for designing bioinformatics algorithms recently described [20]. The MIA method is currently proposed by strictly implementing the idea of MPA and specially designed for variable selection of support vector machines. It works by first computing a large number of SVM classifiers using randomly sampled variables. Each model is associated with a margin. Then, the nonparametric Mann-Whitney U test [21] is employed to calculate a p-value for each variable, aiming at uncovering the variable that can increase the margin of a SVM model significantly. The rationale behind MIA is that the performance of SVM depends heavily on the margin of the classifier. As is known, the larger the margin is, the better the prediction performance will be. For this reason, variables that can increase the margin of SVM classifiers should be regarded as informative variables or possible biomarker candidates. On the whole, the main contributions of MIA are two folds. First, it is originally from model population analysis which helps statistically establish variable rank by analyzing the empirical distributions of margins of related SVM classifiers. Second, it explicitly utilizes the influence of each variable on the margin for variable selection. The results for two publicly available microarray data sets show that MIA typically selects a small number of margin-influencing informative genes, leading to comparable classification accuracy compared to that reported in the literature.

2 Theory and Methods

2.1 Support Vector Machines with Linear Kernel

Support vector machines, based on margin maximization, is a promising kernel-based method for data mining and knowledge discovery [15], [22], [23], [24]. It stems from the framework of statistical learning theory or Vapnik-Chervonenkis (VC) theory and was originally developed for pattern recognition problems. In the present study, we focus on SVM with linear kernel. Such a model is easy to interpret and can hence help understand the mechanism that underlies the data. The theory of linear SVM is briefly introduced in the following.

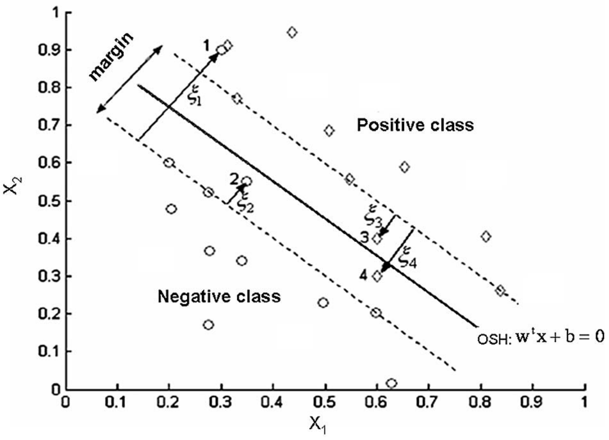


Fig. 1. Slack variables and the optimized separating hyperplane (OSH) for the linearly inseparable data. The distance between the two paralleled dashed line is the so-called margin.

Fig. 1 shows a situation where the two classes of data (diamond and circle) are linearly inseparable. In order to cope with this kind of data, Cortes and Vapnik introduced the slack variable to construct the operating separating hyperplane (OSH) by taking into account the inevitable measured errors in data. Assume that the each sample is denoted by $\mathbf{x}_i$ accompanied with a class label y_i (1 or -1), $i = 1, 2, \ldots, m$. The slack variable associated with each sample is ξ_i. Then the constraint inequality for computing SVM models can be expressed in the following form:

$$(\mathbf{w}^t\mathbf{x}_i + b)y_i \ge 1 - \xi_i, \xi_i \ge 0,\ i = 1, 2, \ldots, m, \tag{1}$$

where $\mathbf{w}$ is the weight vector and b is the intercept of a linear SVM model. The margin of support vector machines is defined as the distance between the two dashed paralleled lines (Fig. 1) and can be computed using the following formula:

$$\mathrm{m\ arg\ in} = \frac{2}{||\mathbf{w}||}. \tag{2}$$

By maximizing the margin, the computation of a SVM model can be formulated as the following optimization problem:

$$\begin{aligned} &\mathrm{min\ imize:}\ \frac{1}{2}||\mathbf{w}|| + \mathbf{C}\sum_{i}^{m}\xi_i, \\ &\mathrm{subject\ to:}\ (\mathbf{w}^t\mathbf{x}_i + b)y_i \ge 1 - \xi_i, \xi_i \ge 0_i, \end{aligned} \tag{3}$$

where C is a predefined penalizing factor controlling the trade-off between the training error and the margin. By using quadratic programming (QP) algorithm, the linear SVM classifier can be computed and expressed as

$$f(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}^t\mathbf{x} + b) = \mathrm{sgn}\left[\left(\sum_{i=1}^{m} y_i\alpha_i\mathbf{x}_i\right)^t \mathbf{x} + b\right]. \tag{4}$$

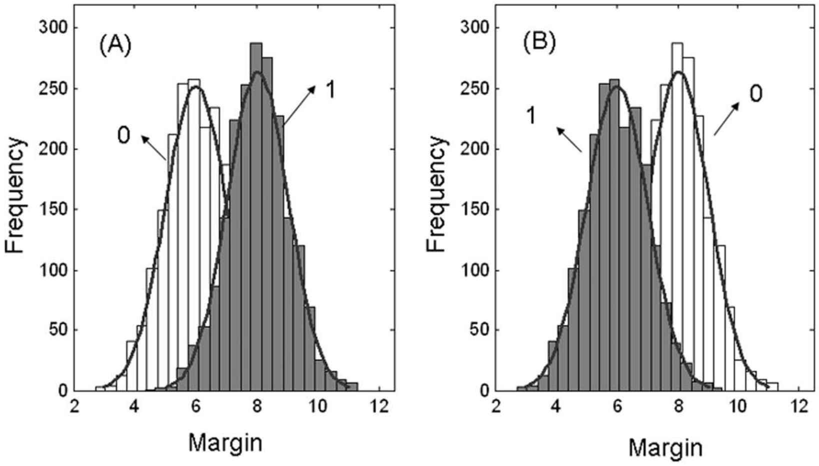


Fig. 2. The two types of variables obeying different margin distributions. Plot A: variables with $\mathrm{DMEAN} > 0$. Plot B: variables with $\mathrm{DMEAN} < 0$. The peak denoted by “1” stands for of models including a given variable, while the peak denoted by “0” is the margin’s distribution of models not including the given variable.

For more details on SVM, readers are referred to several tutorials [15], [24], [25].

2.2 Margin Influence Analysis for Variable Selection Based on MPA

As mentioned before, MPA refers to the analysis of a large number of submodels [20]. It can be conducted in three successive steps: 1) Obtain N subdata sets by Monte Carlo sampling (MCS), 2) Establish a submodel for each subdata set and 3) Statistically analyze some interesting outputs of all the N submodels. The key point of MPA is how to conduct statistical analysis of the interesting outputs, e.g., margins of SVM, of all the submodels for achieving some special goal, e.g., outlier detection or variable selection. Details on MPA could be found in our previous work [20]. In this section, the margin influence analysis is developed by strictly implementing the idea of MPA.

2.2.1 Monte Carlo Sampling in the Variable Space

Suppose that we are given a data set ($\mathbf{X}$, $\mathbf{y}$) consisting of m samples in the rows and p variables in the columns. The class label vector y is of size $\mathrm{m} \times 1$, with element equal to 1 or -1 for the binary classification case. The number of MCS is denoted by N (usually large, e.g., 10,000). With such a setting, Monte Carlo sampling in the variable space can be conducted in three steps: 1) predefine the number of variables, denoted by Q, to be sampled, 2) in each sampling, randomly pick out without replacement Q variables from among the p variables thus obtaining a subdata set of size $\mathrm{m} \times \mathrm{Q}$. Repeat this procedure N times, and we can get N subdata sets. All the sampled N subdata sets are denoted as $(\mathbf{X}\mathrm{sub}, \mathbf{y}\mathrm{sub})_{\mathrm{i}}, i = 1, 2, 3, \ldots, N$.

2.2.2 Submodel Building Using SVM

Given a penalizing factor C, one can build a linear kernel-based SVM classifier for each of the randomly sampled subdata sets. In the current work, C is chosen by cross validation [26], [27], [28]. Therefore, N SVM classifiers together with N margins can be computed. The N margins are denoted by $M_i, i = 1, 2, \ldots, N$.

2.2.3 Statistical Analysis of Margin Distribution by Nonparametric Test

In this section, the procedure for uncovering informative variables is established based on the N margins of the N constructed SVM classifiers. Without loss of generality, we take the $i\mathrm{th}$ variable as a case to illustrate the computing procedure.

First, all the N computed SVM classifiers are assigned to two groups, named Group A and Group B. Group A collects all the models which include the $i\mathrm{th}$ variable, while Group B collects all the models which do not include this variable. Assume that the number of models in Group A and B are $\mathrm{N}_{\mathrm{i,A}}$ and $\mathrm{N}_{\mathrm{i,B}}$, respectively. Then, we have

$$N_{i,A} + N_{i,B} = N. \tag{5}$$

Naturally, we can also get $\mathrm{N}_{\mathrm{i,A}}$ and $\mathrm{N}_{\mathrm{i,B}}$ margins associated with SVM classifiers in Group A and Group B, respectively. Further, we can compute two distributions corresponding to the $N_{i,A}$ and $N_{i,B}$ margins, respectively. Denote the mean values of the two distributions by $\mathrm{MEAN}_{\mathrm{i,A}}$ and $\mathrm{MEAN}_{\mathrm{i,B}}$, respectively. The difference of the two mean values can written

$$\mathrm{DMEAN}_i = \mathrm{MEAN}_{i,A} - \mathrm{MEAN}_{i,B}. \tag{6}$$

From (4), one expects that the inclusion of the $i\mathrm{th}$ variable in a model increases the margin if $DMEAN_i > 0$. In the present study this type of variable is treated as candidates of informative variables. In contrast, if $\mathrm{DMEAN}_i < 0$, one may infer that including this variable into a model will decrease the margin of the SVM models and thus reduce the predictive performance of the model. By analogy, variables of this type are called uninformative variables. The two kinds of variables are illustrated in Fig. 2. Plot A and Plot B show the introduced two types of variables, respectively.

After deriving the margin’s distribution of each variable, we proceed to identify the informative variables in three successive steps: 1) remove all the variables with $\mathrm{DMEAN}_{\mathrm{i}} < 0$, 2) use Mann-Whitney U test [21] to check

TABLE 1
Simulation Study: Means of Classification Error (with Their Standard Errors in Parentheses Computed by Running MIA on 200 Randomly Produced Data Sets for Each Setting)*

(n_H, n_D)	π = 0.05		π = 0.5	
	small change	large change	small change	large change
(15, 15)	0.11±0.11	0.01±0.03	0.15±0.12	0.01±0.03
(20, 10)	0.13±0.11	0.01±0.03	0.14±0.11	0.01±0.03
(50, 50)	0.05±0.05	0.01±0.01	0.05±0.05	0.01±0.01
(70, 30)	0.05±0.05	0.01±0.01	0.05±0.05	0.01±0.01

*: The Q values for π = 0.05 and 0.5 are set to 50 and 200, respectively. The number of Monte Carlo samplings is fixed at 5000.

whether the increment of margin is significant, leading to a p value for each variable and 3) rank the variables using the p value. In this sense, the variables with p value smaller than a predefined threshold, e.g., 0.05, are defined as informative variables in this work. The informative variables should be treated as the most possible biomarker candidate. The margin's distributions of informative and uninformative variable are illustrated in Fig. 2. It should be noted that the proposed MIA method can also be applied to SVM with nonlinear kernels, e.g., Gaussian kernel because only the distribution of margins is required.

3 Numerical Experiments

3.1 Simulation Study

We use the same simulation settings as originally described by Ghosh and Chinnaiyan [11]. We consider the following sample size combinations $(\mathrm{n_H}, \mathrm{n_D}) = (15, 15), (20, 10)$, $(50, 50)$, and (70, 30), where $\mathrm{n_H}$ and $\mathrm{n_D}$ denote the number of samples in healthy group and case group, respectively. Each sample is generated as a vector of 1,000 variables/genes in which a fraction π of the genes was differentially expressed between the two classes. $\pi = 0.05$ and $\pi = 0.5$ were considered in this study. These settings have also been studied by Ma and Huang [10]. For each simulated data set, 2/3 of the samples are randomly selected as the training set and the remaining 1/3 of the samples used as the test set. For each setting, 200 simulated data sets were randomly generated. The Prediction errors in terms of mean and standard deviation are given in Table 1.

It was found that in all simulated settings, the prediction errors based on the variables selected by MIA are satisfactory. Compared to the results by Ghosh et al. (2005, Table 1), our results are much better except for the two settings: $(\mathrm{n_H}, \mathrm{n_D}) = (15, 15)$ and (20, 10) with small change when $\pi = 0.5$. Compared to those results of Ma and Huang (2005, Table 1), one can also find that MIA achieves lower misclassification rate except for the four settings: $(n_H, n_D) = (15, 15)$ and (20, 10) with small change when $\pi = 0.05$ and 0.5. These results indicate that MIA is a good alternative for variable selection of high dimensional data.

We have also tested the computational cost of MIA and got positive results. Detailed information can be found in the supplementary materials (TimeCost.pdf, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36). Also, it is shown that MIA is, to some degree, robust to noisy variables (data not shown).

3.2 Colon Data and Estrogenc Data

3.2.1 Data Description

The original colon data set contains the expression profiles of 6,500 human genes measured on 40 tumor and 22 normal colon tissues by applying the Affymetrix gene chip technology. A subset of 2,000 genes with the highest minimal intensity across the samples have been screened out by Alon et al. [29] and are also made publicly available at http://microarray.princeton.edu/oncology/. The estrogen data set was first reported by West et al. [30] (2001) and by Spang et al. [31]. It consists of the expression values of 7,129 genes measured on 49 breast tumor samples. Of these samples, 25 samples are LN positive and the remaining 24 ones are LN negative. The raw data are publicly available at http://mgm.duke.edu/genome/dna_micro/work/. Before gene selection and classifier building, the data were pretreated using the same methods as described by Ma and Huang [10], resulting in 3,333 genes for further analysis.

3.2.2 Tuning Parameter Selection and Model Validation

As discussed in Section 2, there are a total of three tuning parameters in the MIA algorithm. They are C, the penalizing factor of SVM, Q, the number of sampled variables for drawing subdata set and N, the number of Monte Carlo samplings. For both data sets, C was chosen by cross validation. Concerning the choice of Q, we examined the reproducibility of the identified informative variables by MIA and the corresponding prediction errors with Q set to 20, 50, 100, and 200, respectively. The results for both data sets are shown in the supplementary material (see Table S1 and Qcompare.pdf, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36). For both data sets Q equal to 200 was found to give low prediction error and standard deviation. As for the number of Monte Carlo samplings, the larger N should give better results but at higher computational cost. Considering the computational cost and also the reproducibility (see Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36), we choose $N = 10{,}000$ in the present work. Before running MIA, each gene was standardized to zero mean and unit variance across all the samples. Since the number of samples was small, the leave-one-out cross validation (LOOCV) based classification error was used to validate the performance of the selected genes. This is in line with established procedures in the literature [3], [32].

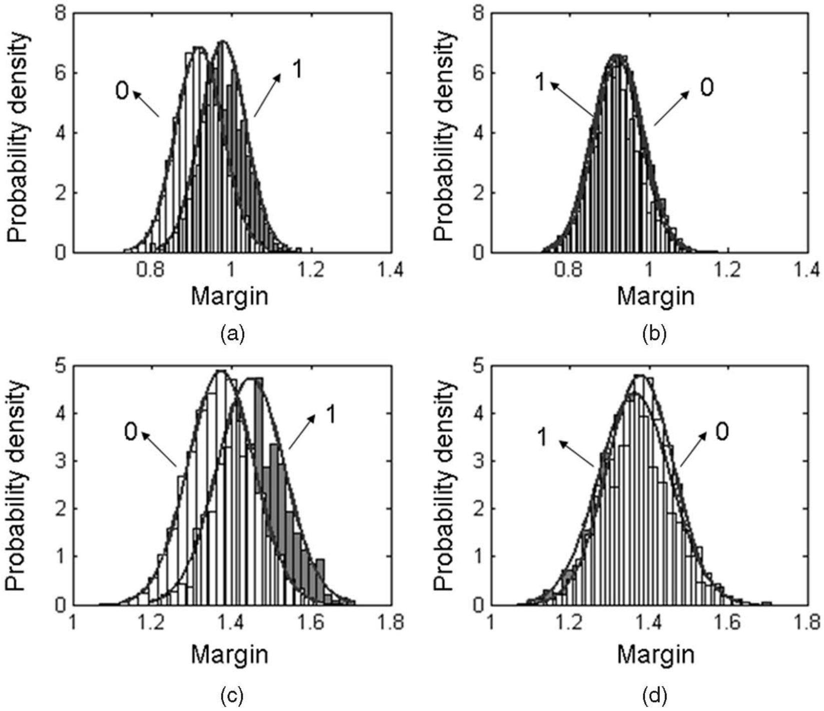


Fig. 3. The illustration of both informative and uninformative genes by means of the proposed margin distribution. Plots A and C show the margin distribution of two informative genes for colon and estrogen data, respectively. By contrast, Plots B and D show the margin distribution of two uninformative genes for colon and estrogen data, respectively.

3.2.3 Results and Discussion

For the colon data, 1,219 out of the 2,000 genes were identified as uninformative genes which decrease the margin of SVM classifiers. After removing these genes, the nonparametric Mann-Whitney U test is applied to test whether the remained 781 genes could significantly increase the margin of the SVM classifiers, leading to a p value associated with each gene. In all, 217 out of the 781 genes were found to be informative with $p \leq 0.05$. Further, in order to control the family-wise error rate (FWER), the Holm-Bonferroni method was utilized to perform multiple testing correction, resulting in 108 significant genes ($\mathrm{p} \leq 0.05$). All the 217 informative genes together with the p values and corrected p values are listed in Table S2 in the supplementary materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36. By using the same procedure, we identified 334 informative genes for estrogen data, among which 108 genes are significant ($p \leq 0.05$) after multiple testing correction. The p values as well as the corrected p values are presented in Table S3 in the supplementary materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2011.36.

As described before, the margin distribution is the basis of MIA for variable selection. Therefore, the margin distribution of an informative gene as well as that of an uninformative one for both data sets are presented here. For colon data, they are shown in Plots A ($\mathrm{Gene\ ID} = 1{,}482, \mathrm{p} = 5.64 \times 10^{-181}$) and B ($\mathrm{GeneID} = 1{,}781$) in Fig. 3, respectively. It is clear that the margin distribution when including the 1,482th gene is right shifted. This means that this gene has the potential to increase the margin of SVM classifiers and can hence improve the generalization performance if included in an SVM model. By contrast, the 1,781th gene decreases the margin and therefore should be removed from the model. For estrogen data, the margin distribution of an informative gene and an uninformative one is shown in Plots C ($\mathrm{Gene\ ID} = 132, \mathrm{p} = 2.19 \times 10^{-85}$) and D ($\mathrm{Gene\ ID} = 1{,}984$) in Fig. 3, respectively. By comparison, it could be observed that the 132th gene can significantly increase the margin of SVM classifiers, whereas the 1,984th gene can only decrease the margin. The above analysis indicates that informative genes can be statistically identified by testing the difference of the interesting parameter, i.e., margin for SVM, when a gene is included or excluded in a model.

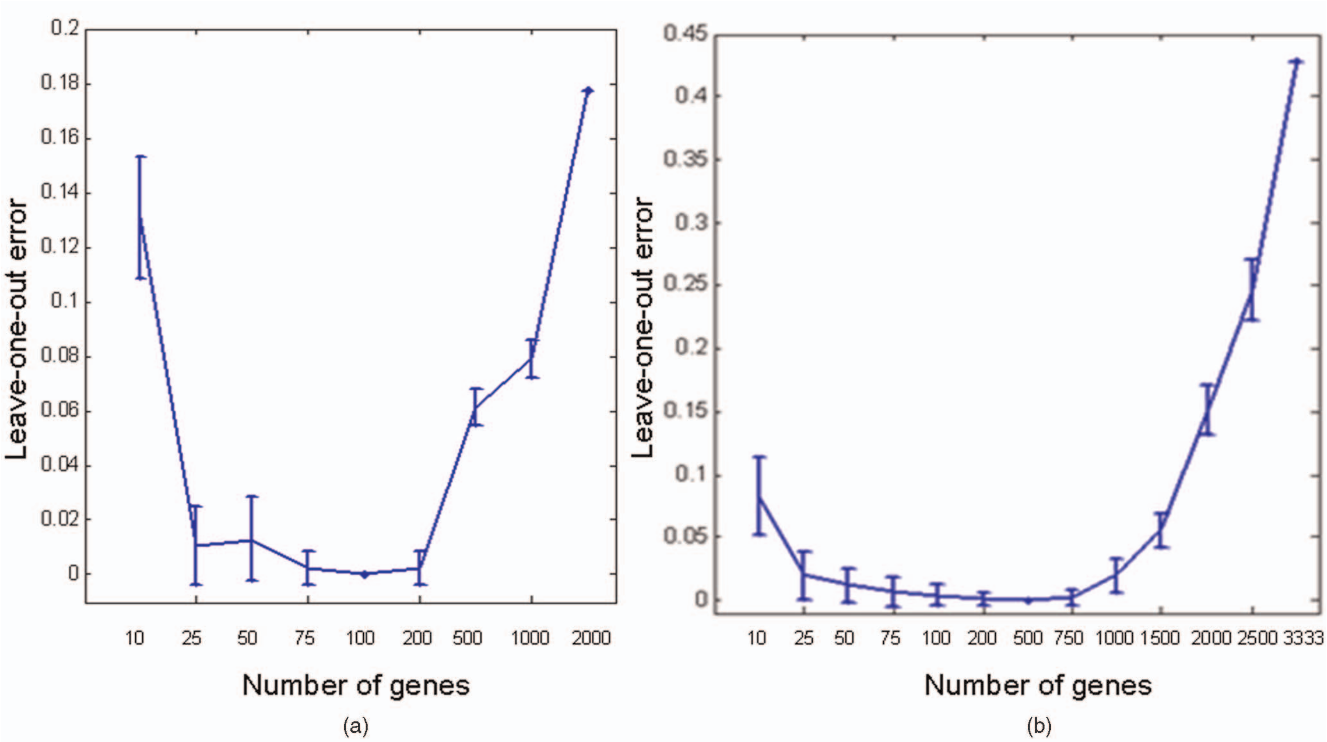
To build a classification model for cancer prediction, a subset of genes should be first identified. Here, we first rank the genes ($\mathrm{DMEAN} > 0$) using the p value. For colon data, nine different gene sets are investigated here. The numbers of the nine gene sets are 10, 25, 50, 75, 100, 200, 500, 1,000, and 2,000, respectively. For estrogen data, 13 different gene sets are considered, of which the numbers of genes are 10, 25, 50, 75, 100, 200, 500, 750, 1,000, 1,500, 2,000, 2,500, and 3,333. Note that the prediction error of the established SVM classifiers could not be exactly reproduced due to the embedded Monte Carlo strategy of MIA. Therefore, we investigated the variation of the classification error in dependence of the number of genes by running MIA procedure on both data sets 20 times. The top five ranked genes of colon and estrogen data are listed in Table 2, respectively. The mean LOOCV errors as well as the standard deviations on colon and estrogen data are shown in. Figs. 4a and 4b, respectively. From Fig. 4c, it can be found that both the mean LOOCV errors and the standard deviation first gradually decrease and then achieve the minimum when 100 significant genes are included. For the estrogen data, it is clear that after including 100 genes, both

TABLE 2
The Top Ranked Five Genes for Colon and Estrogen Data

Data set	GeneID	Gene description
Colon	Hsa.12241	Acetylcholine receptor protein, delta chain precursor(Xenopus laevis)
	Hsa.41098	Human Mullerian inhibiting substance gene, complete cds
	Hsa.10909	Map Kinase phosphatese-1 (Homo sapiens)
	Hsa.36689	H.sapiens mRNA for GCAP-II/uroguanylin precursor
	Hsa.404	Human MXI1 mRNA, complete cds
Estrogen	AFFX-CreX-3_st	Bacteriophage P1 cre recombinase protein
	Z22536_at	Homo sapiens ALK-4 mRNA, complete CDS
	AFFX-BioB-3_at	E coli bioB gene biotin synthetase
	Y10871_at	H.sapiens twist gene
	U37408_at	Human CtBP mRNA, complete cds

mean errors and the standard deviations do not change significantly and achieve the minimum at 500 genes. For both data, the results after gene selection are obviously improved compared to that using all genes, indicating that gene selection is very necessary for improving the prediction ability and the identified informative genes by MIA are actually predictive.

For comparison, the results on both data sets from MIA together with those reported in the literature are listed in Table 3. For the colon data, the minimal classification error from Dettling and Buhlmann [33] was 14.52 percent by using LogitBoost. In Nguyen and Rocke's work [3], the lowest error achieved was 6.45 percent by using PLS-LD. Sigmoid maximum rank correlation (SMRC) was utilized by Huang and Ma, leading to the mean classification error 14 percent



(a)

(b)

Fig. 4. The mean leave-one-out cross validated classification error as well as the standard deviations of 20 runs of MIA. A: colon data. B: estrogen data.

TABLE 3
Comparison of the Leave-One-Out Cross Validation Error by Using Different (Variable Selection) Methods*

Colon		10	25	50	75	100	200	500	1000	2000
A	LogitBoost, optimal	14.52	16.13	16.13	16.13	16.13	14.52	---	---	12.90
	LogitBoost, estimated	22.58	19.35	22.58	20.97	22.58	19.35	---	---	19.35
	LogitBoost, 100 iterations	14.52	22.58	22.58	19.35	17.74	16.13	---	---	16.13
B	PLS+LD	---	---	6.45	---	6.45	---	9.68	8.06	---
	PLS+QDA	---	---	8.06	---	9.68	---	8.06	9.68	---
C	SMRC	---	---	---	---	---	---	14.00	---	---
D		---	---	---	---	---	---	---	9.68	9.68
E	RFE	0.00	3.23	0.00	3.23	4.84	4.84	11.29	9.68	17.74
F	SFS-motivated method	16.13	1.61	0.00	0.00	0.00	0.00	4.84	6.45	17.74
G	MIA	13.06	1.05	1.29	0.24	0.00	0.24	6.13	7.90	17.74
Estrogen		10	25	50	75	100	200	500	3333	7129
A	LogitBoost, optimal	4.08	4.08	2.04	2.04	2.04	4.08	---	---	2.04
	LogitBoost, estimated	6.12	6.12	6.12	6.12	6.12	6.12	---	---	6.12
	LogitBoost, 100 iterations	8.16	6.12	6.12	4.08	4.08	8.16	---	---	6.12
C	SMRC	---	---	---	---	---	---	6.00	---	---
E	RFE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	42.86	---
F	SFS-motivated method	12.24	2.04	0.00	0.00	0.00	0.00	0.00	42.86	---
G	MIA	8.27	1.94	1.12	0.71	0.41	0.10	0.00	42.86	---

*: The results listed here are based on the methods:
A: Marcel Dettling and Peter Buhlmann (2003) [33].
B: D. Nguyen and D.M. Rocke (2002) [3].
C: Ma and Huang (2005) [10].
D: Furey et al (2000) [32].
E: Guyon et al (2002) [17].
F: Gualdrón, et al (2007) [18].
G: The proposed method.

with a standard deviation 7 percent. By using support vector machines, Furey et al. [32] misclassified six samples, resulting in a LOOCV error 9.68 percent. Besides, we have also encoded two variable selection methods: recursive feature selection [17] and sequential forward selection (SFS)-motivated method [18] and performed variable selection on colon data. The LOOCV errors for different numbers of genes are also listed in Table 3. By comparison, it could be found that the proposed MIA is very competitive in gene selection for predicting the colon cancer. For estrogen data, the reported results in the literature are collected in Table 3. The results by using RFE as well as the SFS-motivated method at different number of genes are also presented. On the whole, it might be concluded that MIA is a good alternative for gene selection and the MIA-based SVM classifier is very predictive of the clinical outcome.

4 Conclusions

Based on model population analysis, a new method, margin influence analysis, is proposed to specifically conduct variable selection for support vector machines. With the aid of a "population" of SVM classifiers, MIA has the potential to identify informative variables by statistically analyzing the distribution of margin associated with each gene with the help of Mann-Whitney U test. In this way, one can distinguish the informative variables from the uninformative genes in an easy and elegant manner. Using two publicly available cancerous microarray data sets, it is demonstrated that MIA typically selects a small number of margin-influencing genes and achieves competitive classification accuracy compared to that in the reported literature. The distinguished features and outstanding performance should make MIA a good alternative for gene selection of high dimensional microarray data using support vector machines. It's expected that MIA will find more applications in other fields, such as proteomics and metabolomics.

Acknowledgments

This work was financially supported by the National Nature Foundation Committee of P.R. China (Grants No. 20875104 and No. 10771217), the international cooperation project on traditional Chinese medicines of the Ministry of Science and Technology of China (Grant No. 2007DFA40680), and the Graduate degree thesis Innovation Foundation of Central South University (CX2010B057). The studies meet with the approval of the university's review board. The study is approved by the review board of Central South University. Professor Rolf Manne in Bergen University of Norway is acknowledged for his revisions of our manuscript, especially the language. Yi-Zeng Liang is the corresponding author.

REFERENCES

[1] W. Ma et al., "Support Vector Machine and the Heuristic Method to Predict the Solubility of Hydrocarbons in Electrolyte," *J. Physical Chemistry A,* vol. 109, no. 15, pp. 3485-3492, 2005, DOI doi:10.1021/jp0501446.

[2] T. Hastie, R. Tibshirani, D. Botstein, and P. Brown, "Supervised Harvesting of Expression Trees," *Genome Biology,* vol. 2, pp. research0003.0001-0003.0012, 2001.

[3] D. Nguyen and D.M. Rocke, "Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data," *Bioinformatics,* vol. 18, pp. 39-50, 2002.

[4] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.,* vol. 97, pp. 77-87, 2002.

[5] Y. Lee and C. Lee, "Classification of Multiple Cancer Types by Multi-Category Support Vector Machines Using Gene Expression Data," Technical Report 1051, Dept. of Statistics, Univ. of Wisconsin, Madison, WI, 2002.

[6] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *J. Royal Statistical Soc. B,* vol. 67, pp. 301-320, 2005.

[7] E. Candes and T. Tao, "The Dantzig Selector: Statistical Estimation when p Is Much Larger than n," *Annals of Statistics,* vol. 35, no. 6, pp. 2313-2351, 2007.

[8] Y. Lai, "On the Identification of Differentially Expressed Genes: Improving the Generalized F-Statistics for Affymetrix Microarray Gene Expression Data," *Computational Biology Chemistry,* vol. 30, no. 5, pp. 321-326, 2006.

[9] T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* vol. 286, no. 5439, pp. 531-537, 1999, DOI 10.1126/science.286.5439.531.

[10] S. Ma and J. Huang, "Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data," *Bioinformatics,* vol. 21, no. 24, pp. 4356-4362, 2005, DOI 10.1093/bioinformatics/bti724.

[11] D. Ghosh and A.M. Chinnaiyan, "Classification and Selection of Biomarkers in Genomic Data Using Lasso," *J. Biomedicine Biotechnology,* vol. 2, pp. 147-154, 2005.

[12] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Royal Statistical Soc. B,* vol. 58, pp. 267-288, 1996.

[13] X. Liu, A. Krishnan, and A. Mondry, "An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data," *BMC Bioinformatics,* vol. 6, no. 1, article no. 76, 2005.

[14] W.S. Noble, "What is a Support Vector Machine?," *Nature Biotechnology,* vol. 24, pp. 1565-1567, 2006.

[15] H.-D. Li, Y.-Z. Liang, and Q.-S. Xu, "Support Vector Machines and Its Applications in Chemistry," *Chemometrics and Intelligent Laboratory Systems,* vol. 95, pp. 188 -198, 2009.

[16] Y. Aksu, D.J. Miller, G. Kesidis, and Q.X. Yang, "Margin-Maximizing Feature Elimination Methods for Linear and Nonlinear Kernel-Based Discriminant Functions," *IEEE Trans. Image Processing Neural Networks,* vol. 21, no. 5, pp. 701-717, May 2010.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, no. 1, pp. 389-422, 2002.

[18] O. Gualdrón et al., "Variable Selection for Support Vector Machine Based Multisensor Systems," *Sensors Actuators B-Chemical,* vol. 122, no. 1, pp. 259-268, 2007.

[19] Y. Aksu, D.J. Miller, and G. Kesidis, "Margin-Based Feature Selection Techniques for Support Vector Machine Classification," *Proc. Int'l Assoc. Pattern Recognition (IAPR) Workshop Cognitive Information Processing,* pp. 176-181, 2008.

[20] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, "Model Population Analysis for Variable Selection," *J. Chemometrics,* vol. 24, pp. 418-423, 2010.

[21] H.B. Mann and D.R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other," *Annals of Math. Statistics,* vol. 18, pp. 50-60, 1947.

[22] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[23] V. Vapnik, *The Nature of Statistical Learning Theory,* second ed. Springer, 1999.

[24] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, pp. 121-167, 1998.

[25] K. Hasegawa and K. Funatsu, "Non-Linear Modeling and Chemical Interpretation with Aid of Support Vector Machine and Regression," *Current Computer-Aided Drug Design,* vol. 6, pp. 1-14, 2010.

[26] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. Royal Statistical Soc. B,* vol. 36, pp. 111-147, 1974.

[27] S. Wold, "Cross-Validatory Estimation of the Number of Components in Factor and Principal Component Analysis," *Technometrics,* vol. 20, pp. 397-405, 1978.

[28] Q.-S. Xu and Y.-Z. Liang, "Monte Carlo Cross Validation," *Chemometrics and Intelligent Laboratory Systems,* vol. 56, no. 1, pp. 1-11, 2001.

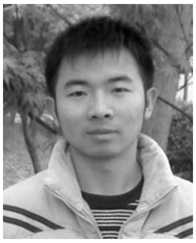
[29] U. Alon et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA,* vol. 96, no. 12, pp. 6745-6750, 1999.

[30] M. West et al., "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 11462 -11467, 2001.

[31] R. Spang, C. Blanchette, H. Zuzan, J. Marks, J. Nevins, and M. West, "Prediction and Uncertainty in the Analysis of Gene Expression Profiles," *Proc. German Conf. Bioinformatics (GCB '01),* 2001.

[32] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics,* vol. 16, no. 10, pp. 906-914, 2000, DOI 10.1093/bioinformatics/16.10.906.

[33] M. Dettling and P. Buhlmann, "Boosting for Tumor Classification with Gene Expression Data," *Bioinformatics,* vol. 19, no. 9, pp. 1061-1069, 2003, DOI 10.1093/bioinformatics/btf867.



Hong-Dong Li received the BSc degree in pharmaceutical engineering in Central South University (CSU) and is currently working toward the PhD degree in CSU. Currently, he is focused on developing variable selection methods for high dimensional data. He proposed model population analysis, which is a general framework for developing advanced bioinformatics methods and proves to be interesting and effective. His current research interests include chemometrics, statistical learning and bioinformatics, and metabolomics.

Yi-Zeng Liang received the PhD degree in analytical chemometrics, Hunan University in 1988. In 1994, he received the Dr. philos. of Norway, University of Bergen. He is a professor of chemometrics and analytical chemistry in Central South University (CSU). He is leading the research centre of modernization of Traditional Chinese Medicines in CSU. He is vice chairman of Computer Chemistry Committee, Chemical Society of China (since 2001), editor of *Chemometrics and Intelligent Laboratory Systems* (since 2007). By far, He has published more than 360 scientific research papers since 1989, in which 280 papers were published in the source journals of SCI and the citation number is more than 4,100 times by SCI source journals with h-index of 29. He has published 10 books (seven in Chinese and three in English) and six chapters (chapter author) in three books in English. His research interests include chemometrics and bioinformatics, chemical fingerprinting, and quality control of traditional Chinese medicines; data mining in chemistry, metabolomics, genomics and proteomics, and analytical chemistry.

Qing-Song Xu received the PhD degree in applied statistics, Hunan University in 2001. From January 1999-April 1999, he conducted the visiting research in Statistics Research and Consultancy Centre of Hong Kong Baptist University. He worked as a postdoctor research fellow in Vrije University of Brussels in Belgium. He is now professor in the school of Mathematical Science and Statistics in Central South University (CSU). His research is mainly focused on chemometrics and bioinformatics. His current research interests include cluster and discriminant analysis, multivariate calibration, nonparametric regression, etc.

Dong-Sheng Cao received the BSc degree in pharmaceutical engineering and the MSc degree in analytical chemistry in Central South University (CSU), and is currently working toward the PhD degree in CSU. His main interests are focused on robust modeling, say, developing outlier detection methods for high dimensional data and ensemble learning. His research interests include chemometrics, statistical learning, and bioinformatics.

Bin-Bin Tan received the MSc degree in pharmaceutical engineering in Central South University (CSU) and is currently working toward the PhD degree in Shanghai Jiaotong University. Her research interests include metabolomics and some chemometrics.

Bai-Chuan Deng received the BSc degree in pharmaceutical engineering in Central South University (CSU) and is currently working toward the PhD degree in Bergen University in Norway. His research interests are focused on chemometrics (especially the resolution of hyphenated data) metabolomics.

Chen-Chen Lin received the BSc degree in biology in Central South University (CSU) and is currently working toward the PhD degree in Bergen University in Norway. Her research interests are focused on chemometrics, metabolomics, and molecular biology.